Agricultural and Forest Meteorology 218–219 (2016) 74–84

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods

Michael D. Johnson[a,*], William W. Hsieh[a], Alex J. Cannon[b,1], Andrew Davidson[c], Frédéric Bédard[d]

[a] Department of Earth, Ocean and Atmospheric Sciences, The University of British Columbia, 2207 Main Mall, Vancouver, BC V6T 1Z4, Canada
[b] Pacific Climate Impacts Consortium, University House 1, PO Box 1700 Stn CSC, University of Victoria, Victoria, BC V8W 2Y2, Canada
[c] Earth Observation, Agriculture and Agri-Food Canada, Science and Technology Branch, 960 Carling Avenue, Ottawa, ON K1A 0C6, Canada
[d] Agriculture Division, Statistics Canada, 170 Tunney's Pasture Driveway Jean-Talon Building, 12th floor D2 Ottawa, ON K1A 0T6, Canada

ARTICLE INFO

Article history:
Received 20 December 2013
Received in revised form 20 October 2015
Accepted 2 November 2015

Keywords:
Canadian Prairies
Crop yield forecasting
Remote sensing
MODIS
NDVI
EVI
Machine learning
Artificial neural networks

ABSTRACT

Crop yield forecast models for barley, canola and spring wheat grown on the Canadian Prairies were developed using vegetation indices derived from satellite data and machine learning methods. Hierarchical clustering was used to group the crop yield data from 40 Census Agricultural Regions (CARs) into several larger regions for building the forecast models. The Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) derived from the Moderate-resolution Imaging Spectroradiometer (MODIS), and NDVI derived from the Advanced Very High Resolution Radiometer (AVHRR) were considered as predictors for crop yields. Multiple linear regression (MLR) and two nonlinear machine learning models – Bayesian neural networks (BNN) and model-based recursive partitioning (MOB) – were used to forecast crop yields, with various combinations of MODIS-NDVI, MODIS-EVI and NOAA-NDVI as predictors. Crop yield forecasts made using predictors from July and earlier were evaluated by the cross-validated mean absolute error skill score (in reference to climatological forecasts) during 2000–2011. While MODIS-NDVI was found to be the most effective predictor for all three crops, having MODIS-EVI as an additional predictor enhanced the forecast skills. While MLR, BNN and MOB all showed significantly higher skills than climatological forecasts for all three crops, barley was the only case where the nonlinear BNN and MOB models showed slightly higher skills than MLR. The lack of skill improvement by nonlinear models over MLR is likely due to the short (12 years) record available for MODIS data, which limits our study to 2000–2011, with very low yields coming from a single severe drought year (2002).

1. Introduction

The major field crops from the Canadian Prairies include barley, canola and spring wheat. The production of field crops plays an important role in the economy of the Canadian Prairies and crop yield is a key element for rural development and an indicator for national food security (Li et al., 2007). Early and reliable crop yield forecasts over large areas would help policy makers and grain marketing agencies in planning for exports and imports. The development of forecast models from remotely sensed data is ideal because remote sensing techniques have the potential to provide quantitative and timely information on agricultural crops over large areas and many different methods have been developed to estimate crop conditions (Li et al., 2007).

Crop yield can be influenced by many factors such as precipitation and temperature during growing season (Guo and Xue, 2012; Qian et al., 2009), soil conditions (Alvarez, 2009; Qian et al., 2009), disease, and anthropogenic factors like irrigation or fertilizers (Prasad et al., 2006). Climatic data are often readily available, but some of the other factors can be difficult or impossible to quantify. Remotely sensed vegetation indices can directly measure crop growth and account for many of these factors. The Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) derived from the Moderate-resolution Imaging Spectroradiometer (MODIS) on the Terra satellite have been used to monitor crop conditions in many locations including Canada. Mkhabela et al. (2011) showed that MODIS-NDVI could be used to effectively predict crop yields on the Canadian Prairies one to two months before harvest. Bolton and Friedl (2013) used both MODIS-EVI and NDVI to forecast maize and soybean yields in the Central

* Corresponding author. Tel.: +17788750045.
E-mail address: mjohnson478@mts.net (M.D. Johnson).
[1] Current address: Climate Research Division, Environment and Climate Change Canada.

http://dx.doi.org/10.1016/j.agrformet.2015.11.003
0168-1923/

United States and found EVI to be a more effective predictor than NDVI.

In the past crop forecasts relied on vegetation indices developed from the Advanced Very High Resolution Radiometer (AVHRR) which has a spatial resolution of 1 km. The MODIS sensor has a higher spatial resolution (up to 250 m) than AVHRR and could lead to more accurate forecasts. In addition, MODIS addresses some AVHRR problems (e.g. AVHRR's placement of NIR within water vapor absorption region introduces noise) and MODIS has a finer radiometric sensitivity. Lastly, MODIS has better geolocation accuracy and on board radiometric calibration for scaled reflectances (Trishchenko et al., 2002).

Prasad et al. (2006) developed a linear model that used AVHRR-NDVI along with soil moisture, surface temperature and rainfall to predict corn and soybean yield for Iowa and suggested that higher resolution data could further improve the model. Precipitation and temperature data sets can be difficult to work with because they are often only measured at certain sites. Interpolation or spatial averaging techniques would be required to apply such data sets to crop yield prediction problems (Vincente-Serrano et al., 2003).

The Canadian Prairies cover a very large geographical area that has been divided up into Census Agricultural Regions (CARs) by Statistics Canada. Environmental conditions over this large area can vary greatly and developing a single forecast model for the entire region may not be appropriate. Mkhabela et al. (2011) categorized the CARs based on soil type into three distinct agro-climatic zones including sub-humid, semi-arid and arid and then developed a crop forecast model for each agro-climatic zone. A potential problem of this method is that within a CAR, there will likely be multiple soil types, rendering it difficult to classify the CAR into a specific agro-climatic zone. However, it is possible that a clustering algorithm could be applied to the yield data to better group the CARs into larger regions for the purpose of crop yield forecasting.

Most current crop yield forecasting uses linear statistical models, which cannot capture nonlinear relations in the data. The cropland ecosystem is complex and many of the processes involved are non-linear. This causes problems with traditional statistical models, such as multiple linear regression (Jiang et al., 2004). The results of Mkhabela et al. (2011) show a possible nonlinear relationship between MODIS-NDVI and the yield of barley, canola, field peas and spring wheat. Nonlinear neural networks have been successfully used to predict crop yield using remotely sensed vegetation indices. Li et al. (2007) and Kaul et al. (2005) both used artificial neural networks to forecast corn and soybean yields in the USA. This study attempts to use fully nonlinear models from the field of machine learning, such as model-based recursive partitioning (MOB) (Zeileis et al., 2008) and Bayesian neural networks (BNN) (van Hinsbergen et al., 2009) to forecast crop yields.

This study is focused on improving crop yield forecasting on the Canadian Prairies and therefore the research goals are to: (i) Compare the effectiveness of several vegetation indices derived from satellite data (MODIS-NDVI, MODIS-EVI and AVHRR-NDVI) in forecasting crop yields on the Canadian Prairies and (ii) compare linear and nonlinear statistical/machine learning models to determine which is most appropriate for forecasting crop yields on the Canadian Prairies.

2. Data and methods

2.1. Study area

The Canadian Prairies consists of the provinces of Alberta (AB), Saskatchewan (SK) and Manitoba (MB). Collectively these provinces have about 30 million ha of crop land. The Prairies extend northward from approximately 49°N to 54°N latitudes and westward from approximately 96°W to 114°W longitudes (Mkhabela et al., 2011). Statistics Canada has divided the agricultural land in Canada into Census Agricultural Regions (CARs) (see supplementary materials, Fig. S1). There are 40 CARs in total with 8 in AB, 20 in SK and 12 in MB. The geography and climate of the Canadian Prairies can change from one location to the next. The development of one crop yield forecast model for all of the Canadian Prairies and developing one model for each CAR will likely be suboptimal due to the small number of data points used to train the model. Grouping CARs which behave similarly with respect to yearly crop yield will likely be a better approach. Mkhabela et al. (2011) grouped the CARs based on soil type into three agro-climatic zones: sub-humid, semi-arid and arid because it was thought that crops in each region would behave similarly and this would improve the accuracy of the forecasts. In this study a clustering model for each crop was developed in an attempt to determine if there is a way to group the CARs that will lead to higher forecast skills. A clustering model groups the data into a number of clusters, so that the data in each cluster are more closely related to each other than data from other clusters. The clustering model grouped the CARs based on their crop yield time series over the period 2000–2011 and a forecast model was then developed for each cluster. A problem with clustering analysis encountered in this study was that the optimal number of clusters was not known. To address this issue the hierarchical clustering method was used, which started with each data point in its own cluster and then the data were merged as the number of clusters decreased. The Ward algorithm for computing the distance between clusters was selected (Murtagh and Legendre, 2011). After reviewing the dendrograms (Wilks, 2006) it was estimated that the optimal number of clusters was in the range of 1 to 6 because for each crop the distance between cluster centers appeared to significantly decrease for more than 6 clusters (Fig. S2). Each type of model was developed with data divided into 1 to 6 clusters and the results were compared to determine the optimal number of clusters.

2.2. Crop yield data

The Agricultural Division of Statistics Canada collects detailed annual crop yield data for all crops across Canada. Yield data for barley, canola and spring wheat by CAR (kg ha^{-1}) for the period of 2000–2011 were obtained from Statistics Canada. Some of the yield data were classified as confidential or missing, but for both barley and canola there was less than 0.9% of the data missing and only 4.8% missing for spring wheat, as a result the missing data points were simply ignored. The mean yield value for each crop in each CAR was calculated over the period from 2000 to 2011. As differences in climate, topography, soil type etc. in each CAR affect the mean yield, henceforth our analysis on the yield will be the deviations from the mean (i.e. crop yield minus the mean yield over each CAR).

2.3. Satellite data

Remotely sensed vegetation indices from the AVHRR and the MODIS sensor are essential aspects of this study. As the AVHRR sensor was not originally designed for monitoring vegetation, it suffers from limitations regarding the design of its red and near infrared channels when formulating NDVI (Fensholt and Sandholt, 2005). Two particularly important limitations of the AVHRR are (a) the overlap of the near infrared channel (0.725–1.100 μm) with a region of considerable atmospheric water vapor absorption (0.9–0.98 μm) that can introduce noise to the remotely sensed signal (Huete et al., 2002; Justice et al., 1991); and (b) the relatively